



Department of Energy



National Science Foundation

---

# National Workshop on Advanced Scientific Computing

July 30-31, 1998

---

*Hosted by the National Academy of Sciences*

## FOREWORD

Over the past two decades, computing and networking speeds have increased exponentially, and computational simulation has emerged as a unique and powerful tool to solve complex scientific and engineering problems. Further advances in various areas of information technology present the promise of tools that will allow us to address scientific problems that cannot be modeled or analyzed with current computing technology, that may require large resource investments to investigate experimentally, or that demand high fidelity simulation to extrapolate well beyond current experience. The National Workshop on Advanced Scientific Computing, was jointly organized by the Department of Energy (DOE) and the National Science Foundation (NSF), in order to determine what facilities, capabilities and human resources were required to exploit these advances and to maintain U.S. leadership in science and technology.

The workshop outcomes highlight the enormous opportunities that may revolutionize our approach to scientific research if we fully exploit the opportunities offered by information technology. The findings and recommendations complement those of the recently published interim report of the President's Information Technology Advisory Committee (PITAC). Both forcefully draw attention to the great strength that leadership in computing and information science and technology has brought the Nation, point out serious deficiencies in our current investments in those areas, and call for federal leadership to address those needs.

This report, in conjunction with the interim PITAC report, will serve as the foundation for the development of a national program plan for a federal initiative in Information Technology and Advanced Scientific and Engineering Computation. The DOE and NSF, in conjunction with other science agencies, are building a strategic partnership to establish a science-driven national infrastructure of terascale computing, communications and advanced simulation. This effort is a major priority of the Administration and we are committed to making this major national asset a reality.

Dr. Rita R. Colwell, Director  
National Science Foundation

Dr. Ernest J. Moniz, Under Secretary  
Department of Energy

# National Workshop on Advanced Scientific Computing

*Hosted by the National Academy of Sciences, July 30-31, 1998*

---

## ***PREFACE***

The National Workshop on Advanced Scientific Computation was organized jointly by the Department of Energy and the National Science Foundation, and was hosted by the National Academy of Sciences in Washington D.C. on July 30-31, 1998. More than 200 scientists, engineers, and computing specialists from all parts of the United States participated in this event. This Report contains a summary of the main conclusions and recommendations reached at that Workshop, followed by reports of three Working Groups --- on Science, Technology, and Integration and Partnerships.

*James S. Langer (Chair)*  
*Physics Department*  
*University of California, Santa Barbara*

*Edward Frieman*  
*Scripps Institution of Oceanography*  
*University of California, San Diego*

*Paul Messina*  
*Center for Advanced Computing Research*  
*California Institute of Technology*

*Larry Smarr*  
*National Center for Supercomputing Applications*  
*University of Illinois at Urbana-Champaign*

*November 17, 1998*

# **TABLE OF CONTENTS**

<b>I. SUMMARY AND RECOMMENDATIONS .....</b>	<b>1</b>
<b>II. REPORT OF THE SCIENCE WORKING GROUP .....</b>	<b>4</b>
<b>SCIENTIFIC AND ENGINEERING EXAMPLES .....</b>	<b>5</b>
Discussion.....	8
<b>III. REPORT OF THE TECHNOLOGY WORKING GROUP .....</b>	<b>10</b>
<b>TECHNOLOGICAL ISSUES .....</b>	<b>10</b>
Programmability.....	11
Storage and Data Management .....	12
Algorithms.....	13
Visualization .....	13
Networking.....	14
<b>STRATEGIES FOR ACTION .....</b>	<b>15</b>
Human Resources.....	16
General Issues .....	16
<b>IV. REPORT OF THE WORKING GROUP ON INTEGRATION AND PARTNERSHIPS.....</b>	<b>17</b>
Preferred Option: An Interagency Committee .....	17
Management Structure.....	17
Conclusion.....	18

## ***I. SUMMARY AND RECOMMENDATIONS***

This is a time of extraordinarily rapid advances in science and technology. At the core of these developments is a near-miraculous increase in the power of computers. Computing speeds and capacities have been growing exponentially for over two decades. It is only within the last several years, however, that scientific computation has reached the point where it is on a par with laboratory experiment and mathematical theory as a tool for research in science and engineering. The computer literally is providing a new window through which we can observe the natural world in exquisite detail.

Just as dramatically, it is only within the last few years that the computer-based Internet has emerged as a vehicle for communicating huge amounts of information throughout the world, among laboratories, businesses and homes. These advances in computing and communication can point to nothing less than a profound transformation of the ways in which we gain understanding, make informed decisions, and enable innovation in modern society.

The United States must remain at the forefront of these developments. Although American scientists made the discoveries and inventions that led to the computer, even they could not have anticipated the speed and significance of the latest advances. Accordingly, the recent Workshop, and several DOE and NSF-sponsored workshops that preceded it, were convened in an effort to assess the present situation and to determine what facilities, capabilities and human resources will be needed in order to meet the challenges ahead of us.

The participants at the DOE/NSF Workshop identified both opportunities that urgently need to be captured and challenges that must be met if our nation is to retain its leadership in computer-related science and technology. We offer the following recommendations as a program for immediate action:

- That the US launch a vigorous effort over the next five years to make accessible to the national scientific and engineering communities computing systems whose speed and capacity are approximately one thousand times larger than those now widely available;
- That, concurrently, the US launch a vigorous effort to develop the software, the algorithms, the communication infrastructure, and the visualization systems necessary for effective use of this next generation of computing facilities;
- That the US scientific and engineering communities prepare to use these computing facilities to solve complex problems of both basic and strategic importance; and that the hardware, software and communications developments be coordinated with these scientific and engineering applications;
- That these efforts also serve as the focus of educational initiatives that will prepare young scientists and engineers to participate in novel multidisciplinary projects and, more generally, that will stimulate interest in science among diverse sectors of the US population.
- That these be interagency efforts, carried out jointly by the DOE, the NSF, and potentially other agencies such as NASA, DOD, NIH, and DOC (especially NOAA and NIST) and coordinated at the level of the NSTC;
- That decisions, especially those regarding allocation of resources for scientific projects at the new facilities, be made via an open, peer-reviewed process in order that the highest standards be maintained;
- That target dates for completion of these efforts --- that is, for reaching the point where the new facilities are being used to produce meaningful results --- be the year 2000 for machines at the few-teraflops level and 2003 for machines at the level

of approximately 40 teraflops. Because of the speed at which developments are occurring, there should be a thorough review and possible redirection of these projects in the year 2003.

Such a program would be a highly cost-effective investment of national resources in areas of clear federal responsibility. In addition to maintaining US preeminence in critical fields of research, a visible and skillfully implemented initiative of this kind would have substantial positive impacts on the overall health of science and technology in this country. The scale of these efforts, their relevance to long-term federal missions, and the need for coordination across many sectors of the US scientific and engineering communities, mean that federal leadership will be required for success.

In the paragraphs that follow within this Summary, we add some general remarks about both the direct benefits to be gained from such an initiative and our reasons for believing that its influence will extend well beyond its immediate scientific goals.

The new status of advanced scientific computation, as a unique research tool, on a par with experiment and theory, is the result of its having passed through a threshold of applicability. Scientists now can begin to perform predictive simulations, for example, of atomic-scale behavior in systems containing many millions of molecules, or of detailed fluid flows and chemical reactions in realistically complex engineering applications. Only a very few people even dreamed of such capabilities just a decade ago.

It is essential to recognize that numerical simulation enhances, and does not substitute for, experimental and theoretical research. Meaningful simulations are based on reliable experimental and theoretical inputs, and their outputs are useful only if validated in the laboratory or the manufacturing plant. The best scientific simulations lead to new theoretical understanding which, in turn, leads to experimental discoveries. The best engineering simulations improve the quality of products while greatly reducing the time and cost of moving from concept to market. Moreover, some of the most challenging scientific problems are brought to light in

efforts to develop predictive engineering simulations.

There are three broad categories of problems for which a new generation of advanced numerical techniques will be crucial:

1. Strategic problems for which the underlying scientific principles are thought to be understood, and whose complexity is such that satisfactory solutions are out of range of current computing technology --- but potentially within range of the technology envisioned in this report. For example, the current efforts to predict global climate change and to refine the design of internal combustion engines fall within this category.
2. Strategic problems for which the underlying scientific principles are not well enough understood at present to justify direct simulations, but where the scientific questions themselves are likely to be answered, at least in part, by advanced numerical methods. Examples --- mentioned here solely for purposes of illustration and not as endorsements --- include improving the processing and performance of structural materials, estimating earthquake hazards, or understanding and predicting the behavior of a wide range of biological systems.
3. Fundamental problems where there is reason to believe that a next generation of computational capability will produce major advances. One example is lattice gauge theory in elementary particle physics. Another is formation of structure in the early universe.

In the Report of the Science Working Group, we present a broader set of examples of the scientific and engineering problem areas where advanced numerical techniques are likely to have a major impact. We also point out in that Report that, while some of the most urgent of these problems are coming within reach of the next generation of computing systems, there is now a growing lack of capacity in the US for doing the important computations that already are feasible. The fields where scientific computation is having

the biggest impact are moving rapidly, and there are demonstrated needs for new facilities at all levels. The success of the high-end efforts recommended here will depend on sustained progress in the underlying areas of experimental, theoretical and, especially, computational research.

In the Report of the Technology Group, we discuss computational capabilities that are presently available, capabilities that we expect to gain within the next several years and areas that will require intense efforts in order to achieve the goals stated here. Prominent among the latter areas is software development. The complexity of the architecture of today's high-end systems and their ability to generate massive amounts of data are increasing more rapidly than our ability to use these systems effectively. The Interim Report of the President's Information Technology Advisory Committee (PITAC) emphasizes this point. We believe that our conclusions are consistent with and provide support for the major recommendations of the PITAC Report.

In addition to the technical issues that it raises, the challenge of software development is a problem of human resources. To solve it, we must encourage well-educated and innovative people to enter this field, and we must provide incentives for them to work on projects at the very highest end of the scale of difficulty. There is a serious shortage of such people in the US at present. Similarly, there is a growing shortage of people in the science and engineering disciplines who are able to do cutting-edge, interdisciplinary research in the expanding computational aspects of their fields. These human-resource problems almost certainly can be solved for purposes of achieving the short-term goals recommended here, but they pose far more serious obstacles for the future. An important indirect impact of the effort that we propose, if pursued vigorously and given public attention, ought to be a growth of interest in these fields among US students and young scientists and engineers.

The projects that we are recommending are multi-disciplinary to an extent that is unprecedented in modern experience. Success will require strong interactions between the

interdisciplinary teams that define the scientific and engineering problems, the mathematical scientists who develop the models and computational methods, and the computer scientists who develop the hardware, software, and communications systems. Although the situation is improving, multi-disciplinary coordination of this kind is not naturally consistent with the organization and reward structure at US industrial and governmental laboratories or, especially, at US research universities. Some of the organizational issues associated with multi-disciplinary coordination are discussed in the Report of the Working Group on Integration and Partnerships. The national effort that we propose should focus attention on these issues and provide impetus for much needed change.

It is clear that the changes being brought about by high-end computation and communication go far beyond the practice of science and engineering. Because science and technology are at the core of these developments, a focused effort in advanced scientific computation should be a uniquely effective way in which to make sure that the US maintains leadership in the areas that are moving most rapidly at the moment. But we must be prepared for yet more dramatic changes. Past experience tells us; for example, that today's high-end technologies will become accepted parts of everyday life within a few decades. Yesterday's supercomputers were less powerful than today's laptops; and we are not near the end of this growth process. We can predict with confidence that the new technologies will have important impacts in medicine, in industrial manufacturing, in public safety, and the like. Past experience also tells us, however, that the most important advances will be ones that we cannot now predict.

In summary, the consensus of the participants at the DOE/NSF Workshop is enthusiastic support for the efforts described above, for engaging other agencies to join these efforts, and for encouraging the US scientific and engineering communities to move ahead in these directions with a sense of both urgency and optimism.

## **II. REPORT OF THE SCIENCE WORKING GROUP**

During the past year, both the NSF and the DOE convened working groups to discuss the scientific opportunities and challenges provided by terascale computing. These discussions made it clear that there is a broad set of research areas for which access to a new generation of terascale computers would enable major progress.

These research areas range from investigations of some of the most fundamental questions in science to studies that may have immediate social and economic impacts. They include disciplines that have long made use of large scale simulations -- for example, high energy physics, chemistry, materials, fluid dynamics, aeronautical engineering, atmospheric and ocean sciences, seismology, plasma sciences, astrophysics and general relativity. In addition, there are disciplines for which computing has only recently emerged as a critical tool -- for example, biology, medicine, and library science. The potential impact of terascale computing is very broad indeed.

The science and engineering problems discussed at the Workshop have a number of important points in common. All address issues of great scientific and/or societal importance. These problems are appropriate for high-end computing because their complexity is well matched to the computational capability of leading edge machines. As computing power continues to grow exponentially, it crosses new thresholds of problem complexity, and problems that previously were intractable become routinely solvable.

For instance, solving spherically symmetric problems in fluid dynamics or plasma physics was once at the edge of what could be done with the fastest computers. Spherical symmetry eventually gave way to axial symmetry, then to fully three-dimensional problems, first in steady state and then in dynamic situations. These problems are still typically characterized by single length or time scales. Now we are approaching the ability to solve fully time-dependent three-dimensional problems with multiple length

scales, complex geometries, and a variety of coupled physical processes.

Many fields of investigation are ripe for terascale computing because they are crossing thresholds in the scale and complexity of observational and experimental data. In most of these cases, data acquisition has been funded by federal agencies. The large investments in fields such as astronomy, high energy physics, and the geosciences will not yield their expected returns unless corresponding investments are made in the computational infrastructure that is needed for analysis of the data.

We believe that there is now an unprecedented opportunity to use emerging technologies to create a national information infrastructure for science and engineering. This infrastructure will support remote use of scientific instruments. It will also provide broad access to digital libraries where experimental data sets will be stored, and will support web-based connections to a wide variety of analytic tools.

The growing complexity of the problems being addressed by scientists and engineers is driving the modern trend toward collaborative research. More and more, scientists and engineers need to interact among themselves and also collaborate directly with computer scientists in order to make progress. With the rise of the Internet and the Web, these collaborations no longer require that participants be together at the same place and the same time. There is new flexibility and opportunity, as well as necessity, for productive interactions.

Because computers have become essential tools in so much of modern research, the demand for computational resources exceeds the supply in many areas. The fields where scientific computation is having its strongest impact are moving rapidly, and there are demonstrated needs at all levels. Only the DOE's ASCI program has aggressively tried to put into place people, infrastructure, and programs to address the special computational needs of the Stockpile Stewardship program. In addition to ASCI, however, scientists and

engineers need an advanced computational infrastructure to address many other critical national challenges. In many cases an increase of computing capabilities of one to three orders of magnitude would enable major advances.

## **SCIENTIFIC AND ENGINEERING EXAMPLES**

---

A few examples from the Workshop and Working Group presentations illustrate the broad-ranging needs for advanced computing resources.

- ***Weather and climate prediction*** has always been at the forefront of problems driving the demand for ever-greater computing power. Operational prediction of ten-day weather forecasts was impossible as recently as ten years ago because of the deficiencies in computing capability.

Over the next decade, a new generation of computing power has the potential to make dramatic breakthroughs to extend operational forecasting to predict climate anomalies, such as the El Niño-Southern Oscillation (ENSO), months to years in advance. Further, major advances in computing will enable more accurate simulation and more certain predictions of multi-decade to multi-century global environmental change than is possible today.

Future energy and environmental strategies will require unprecedented accuracy and resolution for understanding how global changes are related to events on regional scales where the impact on people and the environment is the greatest. Achieving such accuracy means bringing the resolution used in weather forecasting to the global predictions, which is not practical currently because of the very large amounts of data storage and long computation times that are required.

A major advance in computing power will enable scientists to incorporate knowledge about the interactions between the oceans, the atmosphere and living

ecosystems, such as swamps, forests, grasslands and the tundra, into the models used to predict long-term change. Climate modeling at the global, regional, and local levels can reduce uncertainties regarding long term climate change, provide input for the formulation of energy and environmental policy, and abate the impact of violent storms.

- ***The human genome project***, building on the ability to decipher the genetic code of living organisms, is providing dramatic insights into how living cells function and has the potential for providing improvements in public health and environmental quality. To realize fully the advances in the genome project and to exploit the advancing flood of DNA sequence data requires advanced computational tools to extract the information contained in DNA sequences.

Providing such tools for modeling and simulation would have profound consequences for the nation. For example, it could change the practice of medicine, giving rise to individualized medicine -- the use of the right drug for the right patient at the right time. Critical steps are to uncover the three-dimensional atomic structure and dynamic behavior of the gene products and to dissect the roles of individual genes and the integrated functions of thousands of genes.

Fully understanding the genetic function requires evaluating the myriad motions that proteins undergo as they act in concert as the cell's machinery. Among other implications, simulations of protein motions contribute to drug design, but pose enormous computational challenges.

The simulation of just one microsecond -- the longest time frame undertaken to date -- of a small protein's life span effectively requires several months on a 256 node Cray T3E computer. Since even very small proteins require tens of microseconds to milliseconds or longer to fold, the computational bottlenecks to answering important biological questions are insurmountable with existing resources and methods. This is a case in

which we clearly need advances in both computation and basic science.

- New space- and ground-based instruments, such as the Hubble Space Telescope, the Keck Telescope, the Sloan Digital Library Survey, and the Cosmic Background Explorer, are creating a revolution in *cosmology* by constructing an increasingly accurate picture of the structure and evolution of the early universe. The goal of physical cosmology is to understand the mechanisms that created this structure, and to discover the physical composition of the universe and the principles that established the initial conditions.

Supercomputers play a major role in this process by enabling cosmologists to simulate model universes for comparison with observations. However, simultaneously to model the universe in the large and galaxy interactions in the small requires computer power far beyond that available today. An increase in computing power by a factor of one hundred over current levels would allow definitive tests of current cosmological models by comparison with the wealth of data being accumulated from satellite and ground-based observations. Such tests are essential to capitalize on the major investments that have been made in the new observational instruments.

- The application of *Einstein's theory of general relativity* to realistic astrophysical processes is bound to bring deep and far-reaching scientific discoveries. It is an essential component for two major directions: high-energy (x-ray, gamma ray) astronomy, and the new frontier of gravitational wave astronomy. The latter promises to provide a new window on the universe through pioneering detectors such as NSF's Laser Interferometer Gravitational-Wave Observatory (LIGO). This new window will provide information about our universe that is difficult or impossible to obtain by traditional observations of electromagnetic radiation.

However, the numerical determination of gravitational waveforms is crucial for

gravitational wave astronomy. The Einstein equations are probably the most complex partial differential equations in all of physics, forming a system of dozens of coupled, nonlinear equations, with thousands of terms in a general coordinate system. With today's supercomputers, we are just crossing the threshold of solving the full Einstein equations for black hole or neutron star collisions, the most powerful sources of gravitational waves known. With an increase in computing power by a factor of about 100, it will become possible to compute a "catalog of gravitational waveforms" to be used as a tool in this new observational astronomy.

- The long-standing goal of *high-energy physics* is to identify the fundamental entities of matter and determine the interactions among them. Remarkable progress has been made towards this goal. We now have fundamental theories of the strong, electromagnetic and weak interactions, which are known collectively as the Standard Model of high-energy physics. The Standard Model has been enormously successful, having passed all experimental tests to which it has been put.

However, it has proven very difficult to extract many of the predictions of quantum chromodynamics (QCD), the theory of the strong nuclear interactions. At present the only means of obtaining the full predictions of QCD from first principles is through large-scale numerical simulations. It is crucial to perform these simulations in order to determine a number of the parameters of the Standard Model, to make precise tests of the Model, to understand the physical phenomena it encompasses, and to determine whether additional theoretical ideas are needed to explain the behavior of fundamental interactions at very high energies or short distances.

QCD simulations are also playing an increasingly important role in support of the very large experimental programs in both high energy and nuclear physics. Definitive QCD calculations typically require one to ten teraflops-years. Thus,

access to terascale computers would enable major advances in our understanding of the fundamental forces of nature.

- Recent major advances in *plasma sciences* have been made in both particle and fluid simulations of fine-scale turbulence and large-scale dynamics, giving increasingly good agreement between experimental observations and computational modeling. Significant innovations have been made in analytic and computational methods for developing reduced descriptions of complex dynamics over widely disparate length and time scales.

For example, in turbulent transport, the full power of the half-teraflops SGI/Cray T3E at NERSC has been used to produce fully three-dimensional, general geometry, nonlinear particle simulations of turbulence suppression by sheared flows. It is important to emphasize that these calculations, which typically used 400 million particles for 5000 time-steps, would not have been possible without access to powerful present generation MPP computers.

Nevertheless, important additional physical features must be included in these as well as other models to produce realistic simulations of plasmas relevant to key applications such as fusion power generation. These more accurate simulations will require the tera- and petascale computational capabilities targeted by the advanced scientific computing program envisioned by the present DOE/NSF Workshop. Indeed, plasma sciences share with many other fields the computational challenge of describing physical processes that span many orders of magnitude in temporal and spatial scales.

- *Seismology and engineering seismology* have been dramatically affected by the emergence of high-performance computing. Simulations in seismology have significant economic and social implications, such as the mitigation of seismic hazards, treaty verification for nuclear weapons, and increased discovery

of economically recoverable petroleum resources.

The problem of computing the ground motion of large sedimentary basins during earthquakes is an example of an important problem in which the physical and mathematical formulations are well understood. However, three-dimensional calculations that involve realistic models and cover the full range of frequencies of interest to structural engineers are not yet possible.

High-end computing resources also are required to advance scientific understanding of seismic wave generation and propagation. Modeling of earthquake rupture processes in the lithosphere represents an important example of this type of problem. Synthetic Aperture Radar and GPS-based sensors are beginning to provide observational data on the scale that is required to address this problem. In most of these areas terascale or even petascale computers will be required.

- *Materials research* combines all aspects of the proposed effort: major potential impact, broad multidisciplinary, and pressing hardware and software issues. The challenge is not only to invent new materials, but also to perfect existing ones by fabrication and processing so that they have the desired performance and environmental response. In other words, we would like computationally to simulate "mature materials" for specific technologies.

Teraflops computation of material evolution during processing and heat-treating could dramatically shorten the time and cost to develop mature materials for high-technology aircraft, automotive, electronic and magnetic storage industries. It is also leading to new regimes such as micro-fabrication processes, nano-scale devices, and semiconductor lasers. Developing mature materials requires teams involving chemists, material scientists and engineers, mechanical engineers, and physicists. Integrative environments will dynamically couple computational

modeling on all lengths and time scales with empirical databases. Computational development of new materials at greatly reduced costs and effort would strongly enhance U.S. industry's competitive edge.

- Accurate simulations of *combustion* systems offers the promise of developing the understanding needed to improve efficiency and reduce emissions by 2010 as mandated by U.S. public policy. Combustion of fossil fuels accounts for 85% of the energy consumed annually in the U.S., and will continue to do so for the foreseeable future.

Achieving predictive simulation of combustion processes will require terascale computing and an unprecedented level of integration among disciplines, including physics, chemistry, mathematics, and computer science. These simulations are expected to lead to new scientific discoveries in chemical reactivity, catalysis, fluid dynamics, and other basic sciences as well as combustion science and combustor design.

Federal investments in experimental combustion science over the last two decades have laid the foundation for a targeted effort in computational models. The general approach that would be used to simulate combustion systems -- coupling engineering simulations of combustion devices to solution of the Navier-Stokes equations for mass and heat transfer and to quantum mechanical resolution of chemical processes -- will stimulate progress in a number of related areas, including chemical manufacturing, pharmaceutical development, and chemically reacting flow manufacturing processes.

### *Discussion*

Participants in the Workshop emphasized that increases in raw processor power must be accompanied by substantial improvements in many other aspects of scientific computation in order for projects of the kinds listed above to be successful. Many of these issues are addressed in more detail in the following Report of the Technology Working Group.

For example, in order to function effectively, the next generation of computers will need greatly increased memory and mass storage capabilities, improved schemes for internal data movement, and advanced input/output devices.

Many Workshop participants indicated that improvements in algorithms and software have played as important roles in advancing their research as have advances in computing power. In order to take advantage of more powerful hardware, substantial efforts in algorithm and software development will be necessary.

Some specific areas that require increased attention include better solution methods, rigorous methods for quantifying uncertainty, integrated problem solving environments, and matching of algorithms to the underlying architectures of terascale machines. For the results to have credibility, development of applications must be accompanied by continuous validation of the physical models being used and of the algorithms, the data, and the overall software systems.

Solving many of the new, increasingly complex problems listed above will require advanced methods in geometry, mesh generation, and data assimilation. There will be a need for new algorithms to take advantage of memory hierarchies and multiple forms of parallelism.

We also must realize that there are many problems for which direct computational approaches are inadequate. Many such problems involve inherently unstable or chaotic physical situations in which small changes in initial conditions lead to large changes in final states. In such cases -- which are extremely common -- there are fundamental limits to what can successfully be computed. Scientists have shown great ingenuity recently in reformulating such problems in ways that allow predictive numerical analysis, but a great deal of fundamental work along these lines remains to be done.

The impact of advanced scientific computation on industry, government, and federal laboratories has been growing for several decades. In the future, advanced

scientific computation will be an indispensable tool for understanding and managing our ever more complex and interrelated world.

In industry, computation will move beyond crash simulations, airplane design, and drug design to a whole new world of data intensive activities such as financial risk management, fraud detection, and supply chain optimization. In government, computing will be extended from its role in national defense to new roles in support of decision making in disaster management, infrastructure maintenance, environmental and energy planning, and the like.

This increase in capability will not occur, however, unless it is accompanied by increased efforts to develop human resources. We need to attract more of our best young people into research on the new applications of the emerging information infrastructure. Government, universities, and industry need to work together to assure broad access to the emerging grid and its computers, distributed datasets, digital libraries, scientific instruments, and scientists. Such a seamless information infrastructure will offer students opportunities to participate in exciting projects that use these new technologies.

If this goal of broad access to the computational grid is realized, then we will accelerate the transfer of these new technologies to the marketplace and to our society at large. The most effective form of technology transfer from academia takes place when young people, trained in the use of new technology, move on from college to the world at large.

In view of the scientific and economic value of the resources arising from the proposed initiative, special attention should be paid to the allocation scheme. (Schemes for allocating resources are discussed also in the Report of the Working Group on Integration and Partnerships.) The priorities of the participating agencies must, of course, be consistent with their missions. Thus, the DOE has decided to focus its efforts on combustion and climate modeling along with selected areas of basic scientific research.

On the other hand, the mission of the NSF is to enable US leadership in all areas of basic research. In order to carry out this mandate, it plans to provide the academic community with broad access to terascale computing facilities. Clearly, there will be considerable overlap between NSF and DOE interests. Therefore, it is important for NSF and DOE to develop new methods for jointly funding areas of mutual responsibility.

The NSF has been very successful in developing processes for the allocation of high end computing resources, first at its individual supercomputer centers, and more recently at the national level through its MetaCenter and National Resource Allocation processes. These processes have provided the stability required for long term projects, while providing the flexibility for new initiatives. They have played an important role in introducing high performance computing into fields that had not previously made use of it. We believe that this working model could be extended to allocate terascale computing in support of NSF and DOE basic research projects.

One of the great strengths of the US system for supporting basic research is the unique way in which individual scientists and engineers through the processes of peer review set scientific directions. These processes include writing and reviewing proposals, and participating in planning workshops and advisory committees. This system has provided agencies such as the NSF with the flexibility needed for changing directions or launching new initiatives as scientific developments warrant.

This flexibility is particularly important in rapidly developing fields such as scientific computation, where it is difficult to predict the most promising areas of research. We recommend that resources for a rigorous peer review process that is open to scientists and engineers in all disciplines allocate research arising from the proposed initiative. If properly coupled to the mission driven investments, a new set of relationships between scientists, engineers and decision-makers will emerge across government, industry, and academia.

### III. REPORT OF THE TECHNOLOGY WORKING GROUP

---

The Technology Working Group addressed two closely related but qualitatively different kinds of questions. The first of these takes a long-range point of view:

*What are the technological issues associated with scaling up to computers with orders of magnitude greater capabilities than those we have now?*

This new generation of computing systems will lead us into technologies that are qualitatively more complex than anything we have seen so far. We can now see that the hardware needed for these systems is feasible and, to some extent, already exists. But we also know that, if these systems are to be used effectively, we shall have to solve a wide variety of challenging problems at the frontiers of the Computer Science. These problems are in the areas of programming, data storage and management, algorithm development, visualization and interpretation of large data sets, networking and, ultimately, integration of all these capabilities into systems that are usable by scientists, engineers, and policy makers. We see the initiative proposed in this Report as impetus toward work in these essential areas.

The second question addressed by this Working Group asks: How do we get there from here?

*What capabilities do we need in order to achieve the specific goals of this initiative? Are those needs realistic? On what time scales? What human resources do we need in order to achieve those goals?*

Our main conclusion is that there are, indeed, feasible strategies for achieving the five-year goals outlined above in the Summary and Recommendations. A balanced portfolio of research and development programs will be necessary. This portfolio must include a full range of efforts extending from university-based individual-investigator and group research

projects to full-scale testbeds for the emerging computational technologies. In many situations, multidisciplinary teams will be required.

This Working-Group report is divided into two main sections addressing, respectively, the long-term technological issues and the more immediate strategies for action.

#### TECHNOLOGICAL ISSUES

---

Unlike the monolithic supercomputers of the 1980s, new high-performance computing systems contain thousands of interconnected microprocessors, many levels of memory, hundreds or thousands of secondary and tertiary storage devices supporting petabyte data archives, and high-resolution visualization systems.

These architectural advances have far outstripped our ability to manage parallelism and to deliver large fractions of peak hardware performance. The complexity of the new systems will require new approaches to programming, compiling, and resource management, as well as new visualization systems and new techniques for data representation.

Moreover, these systems will be interconnected via the emerging national computational grid. They should eventually support multi-language programming, and should be easily used by collaborative research teams based at many different locations. Those teams will need distributed access to very large data archives and sophisticated techniques for information mining and visualization. Building such capabilities will require coordination across a broad range of software specialization's as well as close interactions with the people who are developing algorithms and scientific and engineering applications.

The general consensus among participants at this Workshop was that no simple extension of computing practice as we

know it today will carry us to effective use of the next generation of teraflops systems, i.e. systems requiring thousands of processors to achieve sustained multi-teraflops capability. Rather, we shall need to solve a wide range of technological problems in qualitatively new ways.

This is a daunting prospect. However, the challenges do not all need be met simultaneously. Judicious selection of key technologies for early development should lead to usable systems that then improve over time in functionality and efficiency. A staged approach, in which we do a good job at each level, is much more likely to be successful than an attempt to solve all problems at once. We must not accept mediocre solutions at any level of an advanced computing initiative. Important research advances will be required in every aspect of high-performance computing, and therefore we must put in place a sustained and coordinated, long-term research program.

We turn next to some specific technological issues.

### ***Programmability***

Programming large-scale parallel systems to achieve high efficiency, that is, a high fraction of peak performance, is a major challenge. The fraction of peak performance achieved usually declines with increasing numbers of processors. Often, applications must be carefully matched to the idiosyncrasies of individual systems in order to achieve high performance. As a result, the performance achieved on one parallel system is rarely portable to another system with a different software or architectural substrate.

Two basic high-end computer architectures have emerged in the 1990s from the broad range of parallel architectures developed during the 1980s: distributed shared memory systems and message-based cluster systems. In the former, a large number of processors can directly access a pool of memory modules share by all; in the latter, a relatively small number of processors share memory in each subsystem, and the subsystems access the contents of other

subsystems' memories by sending messages requesting the desired data. Because both architectures contain multilevel memory hierarchies, they both require careful management of data to achieve high performance.

A key programming challenge is developing software that will enable users to maintain high performance while moving code from one kind of platform to the other. Implicit in such an approach is the need for better models of the behaviors of systems and applications. Because both kinds of architecture will co-exist, and because the cluster architectures incorporate distributed memory as well as shared memory features, new programming models are needed.

Current software systems are ill matched to these new computational environments, where transient behavior and competition for shared resources are the norm. To achieve high performance in such environments, applications, libraries, compilers, runtime systems, and system software must dynamically adapt to changing circumstances. Moreover, they must be resilient to hardware and software faults, allowing systems to continue operation when processors or other devices fail, and they must provide naming, security, and authentication for a disparate user base. And somehow, eventually, they must do all these jobs automatically, in ways that are invisible to the user.

Given the complexity of terascale systems, it seems unlikely that fully automated approaches will yield high performance for a broad range of applications in the near future. Instead, semi-automated approaches will assist applications developers, automating certain aspects of resource and data management, as well as code mapping to computational resources. Consequently, application-driven software research must develop approaches that increase automatic exploitation of system capabilities.

We recommend that priority be given to the following areas of concern relevant to programmability:

- Adaptive software for resource management and runtime libraries for heterogeneous assets.
- Dynamically negotiated performance contracts for performance portability.
- End-to-end environments for computation, storage, and visualization.
- Improved reliability and fault tolerance.
- Programming languages suitable for expressing what is to be done with advanced facilities.
- Compiler technology that is capable of dealing with thousands of processors and that takes into account internal processor parallelism as well as system parallelism.
- Instrumentation of early semi-automatic software tools to provide information about critical issues to the compiler-research community.

### ***Storage and Data Management***

Computation is only one aspect of the end-to-end computing environment; data management and user interactions are equally important. Terascale computing systems, together with large-scale scientific instruments, create massive data sets. With multi-terabyte data sets and multi-petabyte data archives come problems of data access, staging, coordination, and transfer. All scientists doing large-scale simulations or analyzing large amounts of experimental data face problems in the areas of data manipulation and storage, visualization and interpretation.

The challenge of making intelligent data manipulation an equal partner with computation highlights the key role of high-performance secondary and tertiary storage systems in both simulation and data mining, and the critical need to exploit parallelism in both computation and data management. Important issues that must be explored include design of intelligent

input/output libraries, techniques for implementing large-scale parallelism across thousands of secondary and tertiary storage devices, support for wide-area access to distributed data archives, and the development of new searching, classification, summarization, and synthesis techniques.

In the commodity storage market, storage densities are rising faster than either rotation speeds are increasing or seek times are decreasing. This means that providing high-speed access to multi-terabyte data sets will require coordination of thousands of storage devices. Moreover, just as programming models must now support deep primary memory hierarchies, terascale storage software must manage tertiary and secondary storage devices with access times ranging from minutes to milliseconds. In turn, this requirement will necessitate creation of new, adaptive, massively parallel systems for searching data files and providing access to small parts of large data sets.

Workshop participants emphasized that increases in raw processing power must be accompanied by corresponding increases in memory and mass storage, by improvements in managing internal data, and by improvements in input/output capabilities. As in the case of programmability, improvements in data storage and management eventually will need to be automated; but we must approach these goals realistically and carefully.

We recommend that priority be given to the following areas of concern relevant to storage and data management:

- Support for multi-terabyte data sets and multi-petabyte archives.
- Adaptive input/output libraries and file systems for concurrent access to thousands of disks and tapes.
- Distributed access to -- and integration of -- multidisciplinary data archives.
- Hierarchical representations for subset extraction and interactive visualization.

- Database and information management systems for intelligent specification of data groupings and correlations.
- “Query by example” interfaces for retrieving items that are similar to one another.
- Cost/benefit analysis for data-storage versus data-regeneration schemes.

### *Algorithms*

Both algorithm development and computer architecture has changed dramatically in the past few years. There have been dramatic improvements in algorithms; particularly those used for solving discretized partial differential equations arising from models of physical phenomena. These improved algorithms, in many useful cases, possess optimal computational efficiency.

For example, the “Multigrid” and “Multipole” methods have implementations that require only a number of order  $N$  operations for approximations involving  $N$  discrete points. Information theory tells us that we can do no better. That is, to solve such problems, every point must be visited at least once; and therefore any algorithm must involve at least a number of operations of order  $N$ . For large-scale problems, the difference in performance between this class of algorithms and earlier ones that scaled as  $N^2$  (or as higher powers of  $N$ ) is often the difference between success and failure of the project.

While those new algorithms were being developed, computer performance was being improved primarily by the use of faster clock rates and more efficient implementation of vector hardware. Now, however, the most powerful computers are parallel systems containing thousands of processors. This change in the picture has presented algorithm designers with an enormous opportunity and challenge.

In this more complex computing environment, it will be even more important than it has been in the past for the specialists in algorithm development to work in close collaboration with, on the one hand, the scientists and engineers whose

applications are being implemented and, on the other hand, with the computer software and hardware experts who understand the capabilities and idiosyncrasies of the emerging systems.

A number of strategies have emerged recently that should be useful in the development of algorithms that will be scalable to thousands of processors. These include adaptive grids, asynchronous algorithms, and load balancing techniques. They also include the use of algorithms to tailor algorithms to be particularly effective for a class of applications; for example, by using optimization techniques to select discrete parameters or using control theory to set continuous parameters.

We recommend that priority be given to the following areas of concern relevant to algorithm development:

- Development of scalable algorithms using strategies such as: optimizing complexity of realistic problems; reducing the size of computations by using unstructured, adaptive grids; and creating asynchronous algorithms with improved latency tolerance and load-balancing capability.
- Development of advanced techniques for design and optimization of algorithms, including the use of design algorithms to optimize applications algorithms.
- Advanced user-architecture interactions to improve performance through techniques such as providing information on cache-memory performance, developing facilities for user control of data movement, and providing feedback to software developers.
- Development of geometry and grid generation methods that deal with as many as one billion cells and that provide adaptability and front tracking.

### *Visualization*

Visualization, virtual environments, and other data presentation techniques are

mechanisms by which we glean insight from both computational and experimental data. They are the primary paths to scientific understanding. However, current visualization systems cannot process and display multi-terabyte data sets – there is a fundamental mismatch between such volumes of data and human perceptual capability. Not only are existing visualization systems unable to present data on such scales; they are not tightly integrated with either parallel computing engines or data archives. Hence, they cannot exploit hierarchical data representations to display, for example, lower resolution imagery in regions of low interest while successively refining the fidelity in regions of high interest.

It is critically important that we be able to present the results of computational simulations and experimental measurements in ways that emphasize important features. Thus, the key research challenge is developing scalable schemes for interactive exploration of large, multi-terabyte data sets. Efficient, interactive exploration of massive data sets will require new compression techniques that automatically segment, cull, extract and summarize relevant features. Moreover, given the scale of multi-terabyte data sets and the need to compare and contrast data from multiple measurements and simulations, scalable visualization systems will require new “intentional” interfaces that allow users to specify intent (e.g., to find other data with features similar to those being observed) rather than merely action (e.g., to move to a specified location).

If we are to reap the benefits of future large-scale computations, these visualization techniques must be connected in ways that will enable scientists to steer simulations, explore outputs, and vary representations of the data interactively. Techniques should include capabilities for on-line “measurement” of data; for example, by “clicking” on a displayed data point to obtain related information. Manipulation of data objects, in the sense of virtual reality, can also be useful. “Flying” through volumetric data is appropriate in some cases, while direct

manipulation of numerical values seems more natural in others.

We recommend that priority be given to the following areas of concern relevant to visualization:

- New scalable architectures for the study of very large data sets.
- New modes of visualization for interpreting the results of large-scale simulations and experiments.
- New human(s)-in-the-loop methods for steering trial computations and monitoring large-scale production simulations.
- New technologies for advanced visualization of data in a variety of physical environments ranging from traditional desktop computers to immersive virtual-reality environments.

## *Networking*

Research in science and engineering increasingly is based on distributed computation across collections of network-connected parallel systems (often called metacomputers or computational grids). In almost every case, scientists are geographically distant from the high-end computing resources that they are using; and even when the scientists and their computers are all at the same site, networks enter the picture.

The computations done in this distributed fashion often are coupled to network accessible scientific instruments, distributed data archives, and real-time visualization and interaction. Thus the user of the computational grids is presented with a highly heterogeneous set of resources. Both the hardware and software capabilities of present networks are being pushed to their limits by these emerging applications.

The advent of terascale computing systems will exacerbate this situation. Moreover, the complexity of the scientific and engineering problems that can be addressed by these systems means that there will be

more and more emphasis on interdisciplinary research, and thus more and more need for advanced networking capabilities.

A particularly relevant example is climate modeling. Here, specialists in several disciplines will need access to oceanic, atmospheric, and satellite data from multiple, geographically distributed archives. They also will want their computers connected in ways that will allow them to steer simulations and vary the representations of their output interactively.

The need for distributed computing, communications, and data access raises new challenges in coordinating end-to-end performance of these extended systems. We shall need to design adaptive, intelligent network software as well as mechanisms for ensuring quality of service. In short, we must create a national information infrastructure that unites disparate data sources, storage and representation formats, and access mechanisms.

We recommend that priority be given to the following areas of concern relevant to networking. (Some of these network capabilities are being developed but may be very expensive. Others will require new research.)

- High bandwidth networks.
- Quality of service.
- Advanced capabilities, such as multicasting.
- High-speed interfaces to connect computers to networks.
- Adaptivity of network software and operating systems to application communication patterns (e.g., ability to change buffer sizes, etc.).
- Adaptivity of applications to network behavior.

- Support for scientific simulations that consist of chains of operations performed by different components on different systems, from reading input data to visualization to archiving results.
- Performance measurement and modeling.

## ***STRATEGIES FOR ACTION***

---

We believe that the following are ambitious but feasible near-term goals for the advanced scientific computing initiative proposed in this Report:

- Acquisition and operation of hardware in the 40-60 teraflops regime by the year 2003.
- Substantial progress in the development of scalable operating systems and software to enable effective use of these facilities.
- Development of new, robust tools for selected applications.
- Design of effective numerical algorithms and libraries for multi-teraflops systems.
- Substantial progress in the development of scalable high-performance data storage, management, visualization, and networking infrastructure.
- Establishment of selected high-performance distributed computing and collaboration environments.

We emphasize that the only way to develop and test large-scale algorithms, software, and tools for data management and visualization is to do so on large-scale systems. Therefore, hundreds of people will need routine access to such systems. An adequate number of full-scale machines will be required, and these will need to be broadly available to scientists and engineers from many disciplines and many different kinds of laboratories.

Are these goals realistic? We believe that they are, especially those pertaining to hardware, if we can take advantage of progress made in ASCI and related programs (e.g., NGI). The goals pertaining to software, data management, and visualization face more serious obstacles, but we are confident that substantial progress can be made. Work on algorithms and development of applications tools needs to be accelerated, but the goals appear to be feasible.

### ***Human Resources***

As is emphasized in many places throughout this Report, one of the most important challenges facing this initiative in advanced scientific computation is the development of human resources. Success will require a broad spectrum of scientists, mathematicians and engineers for research, development, deployment, and support. The supply is limited and there is stiff competition from industry.

Therefore, we need to find ways to fill the human-resource pipeline. Graduate and undergraduate fellowship programs are important components, as are summer internships, and cooperative programs at the high school, undergraduate, and graduate levels. But we also need to stimulate interest in relevant fields; even the supply of computer-science graduates is too small. It will be particularly important that people trained in the physical sciences, who have first-hand experience in solving the kinds of problems for which these advanced computational facilities are being developed, be encouraged to participate in the software and algorithm development aspects of this initiative.

### ***General Issues***

Major research advances in software, data management, and visualization will be required to exploit the promise of terascale computing systems. However, research in each area alone is insufficient. Experience has shown that understanding interactions between components is central to developing robust systems that can meet performance goals. Thus, the committee felt strongly that much of the research must be conducted on realistic testbeds by collaborative teams who shall work closely with applications scientists.

Finally, the Working Group made several general observations:

- The costs, schedules, and the types of people needed for this initiative depend strongly on whether it focuses on commercial production-quality software or on producing specialized tools that will be used by only a few people.
- This is a broad and distributed program that will be difficult to manage, even with adequate human and financial resources.
- Access to terascale computers is as critical for developing the underlying technologies as it is for developing the applications. This requirement is especially important when one realizes that many important problems emerge only in large-scale systems.
- There is a large gap between what the computer industry is prepared to do and what is necessary to develop the truly high-end systems. However, partnerships with industry will be crucial.

## ***IV. REPORT OF THE WORKING GROUP ON INTEGRATION AND PARTNERSHIPS***

The goal of a national effort in advanced scientific computation is to bring the extraordinary new developments in this field to bear on research across a broad range of scientific and technological areas. Because of the scale of these efforts, their relevance to long-term federal missions, and the need for coordination across many sectors of the US scientific and engineering communities, federal leadership will be essential.

The structure of this leadership must be flexible enough to enable changes in the way scientific and technological research is carried out. At the same time, it must be broad enough to accommodate many in both the federal and non-governmental communities who will be drawn to this enterprise. An unusually high degree of coordination across disciplines and between federal agencies, academia, industry, and other users will be necessary. Therefore, a highly adaptive management plan, responsive to the needs for breadth and change, and fostering integration and partnerships must be designed. On the federal side especially, tight coupling of many agencies will be required.

To find a model for interagency cooperation, this Working Group surveyed a number of existing multi-agency programs that are comparable in scope and complexity to the one being considered here. We describe our preferred option in the next paragraphs.

### ***Preferred Option: An Interagency Committee***

Our model for strongly coordinated interagency interactions is based on experiences gained in the early years of the US Global Change Research Program. That program has existed for a decade, has performed outstanding research, and has successfully coped with many interagency issues. The USGCRP budget is about \$2 billion per year.

The specific characteristics to note are:

- Operation as a formal subcommittee of the CENR/NSTC with ten participating agencies.
- Coordination and management of programs, projects, budgets, planning, and execution.
- Successful beginning as a Presidential initiative.
- Strong OMB/OSTP/agency interaction. Funds were "fenced" early on to give the program momentum.
- Ability to fund programs jointly across agencies.
- Strong input from external advisory groups, stakeholders, policy makers, etc.
- Subcommittee responsibility and oversight for programs involving international cooperation.
- External monitoring of overall performance and execution.
- A joint federal project office to facilitate the coordinating functions of the interagency committee.

The general assessment is that the program has prospered and produced outstanding science when the conditions cited above were maintained and acted upon. The early success of USGCRP is one demonstration that a large multi-agency program, created to deal with complex scientific issues, can be adaptive, can create partnerships, and can respond to advice from many components of the scientific community.

### ***Management Structure***

Clearly there is not a one-to-one correspondence between the USGCRP and the present proposal, but there are enough similarities for comparisons to be useful. This

Working Group recommends the following guidelines for management of a new national initiative in advanced scientific computing:

- The program should be tightly coupled under the NSTC umbrella to bring together the many agencies needed to mount a program of this breadth and depth successfully. It is unlikely to achieve its true potential if restricted to a small number of agencies.
- Broad input on a wide variety of issues touching upon this program should be sought in putting together the initial phase of this effort.
- In its early years, this program should be treated as a special Presidential initiative, and fenced funds should be sought for enough time for the program to gain momentum and begin to stand on its own.
- The management structure must support broadly distributed hardware, software, and network capabilities as well as distributed research and application centers. This requirement argues for a system which acquires a great deal of information to keep all informed, but does not too tightly constrain the formation of partnerships and interactions.
- The management should continually seek broad and open input from a wide variety of external sources, and should respond to suggestions and requests in a timely fashion.
- Decisions about participation, support and funding should be made through mechanisms of free and open competition. Joint agency RFP's should be encouraged.
- A government program office for interagency coordination appears to be workable and highly desirable.
- Selections of research and application topics should be made in such a way as to

have optimum impact on the vitality of US basic science and engineering. To assure that this happens, the management group should consist of representatives from both the federal funding agencies and the scientific community at large.

- The program will need time to mature as it adapts to inevitable changes in science and technology, and also as best practices become clear. It is crucial that some guarantees of long-term continuity and stability be provided to permit the true potential to emerge.
- Interagency agreements regarding the responsibilities of the participating agencies and the balance of funding across agencies should be in place as early as possible. The participating agencies must agree to maintain the discipline necessary to keep the program on track to meet its goals.

### *Conclusion*

This working group strongly endorses the goals and basic strategy of the proposed initiative in advanced scientific computation. It agrees with the general results of the Science and Technology Working Groups that set the foundations of the program in their corresponding areas, and finds their arguments about the urgency of the program to be persuasive.

The management issues discussed above, while needing attention and care, are not thought to be impediments to immediate action. Indeed, the Group concludes that something along these lines is likely to succeed and, if organized with sufficient flexibility, will grow and change to accommodate what is needed in the future.

In short, this Working Group concludes that a national effort in advanced scientific computation offers great promise and challenge. It should go forward expeditiously.